

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06290210 A**

(43) Date of publication of application: **18 . 10 . 94**

(51) Int. Cl

G06F 15/38

(21) Application number: **05075638**

(22) Date of filing: **01 . 04 . 93**

(71) Applicant: **SHARP CORP**

(72) Inventor: **MORISHITA TARO
TSUBAKI KAZUHIRO
YAMAJI TAKAHIRO
KOBUCHI YASUJI**

(54) TRANSLATING DEVICE FOR NATURAL LANGUAGE

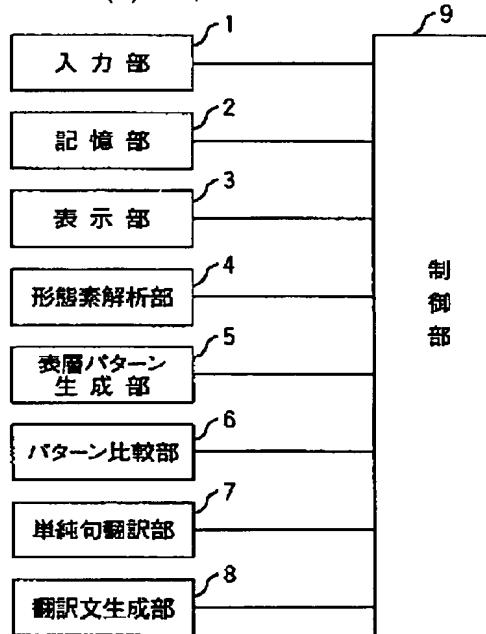
and editing it.

COPYRIGHT: (C)1994,JPO

(57) Abstract:

PURPOSE: To obtain a translation of high quality without analyzing the structure of a sentence and editing it.

CONSTITUTION: A storage part 2 stores a translation example data base. A morpheme analytic part 4 performs a morpheme analysis of an input sentence from an input part 1 to obtain a predicate. A surface layer pattern generation part 5 generates the surface layer pattern of the input sentence on the basis of the morpheme analytic result. A pattern comparison part 6 determines an index of the translation example data base on the basis of the predicate of the input sentence and further retrieves conversion patterns and translation examples under the index by using the surface layer pattern of the input sentence. A simple phrase translation part 7 translates a word string corresponding to category symbols of the surface layer pattern of the input sentence by referring to the translation examples and fill the empty field in the conversion pattern to obtain a character string pattern of a target language. A translation generation part 8 generates a complete translation on the basis of the character string pattern of the target language. Thus, the translation of high quality is obtained without analyzing the structure of the input sentence



(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平6-290210

(43)公開日 平成6年(1994)10月18日

(51)Int.Cl.⁵

G 0 6 F 15/38

識別記号 庁内整理番号

P 7323-5L

E 7323-5L

F I

技術表示箇所

(21)出願番号

特願平5-75638

(22)出願日

平成5年(1993)4月1日

(71)出願人 000005049

シャープ株式会社

大阪府大阪市阿倍野区長池町22番22号

(72)発明者 森下 太朗

大阪府大阪市阿倍野区長池町22番22号 シ
ャープ株式会社内

(72)発明者 樋 和弘

大阪府大阪市阿倍野区長池町22番22号 シ
ャープ株式会社内

(72)発明者 山路 孝浩

大阪府大阪市阿倍野区長池町22番22号 シ
ャープ株式会社内

(74)代理人 弁理士 青山 葵 (外1名)

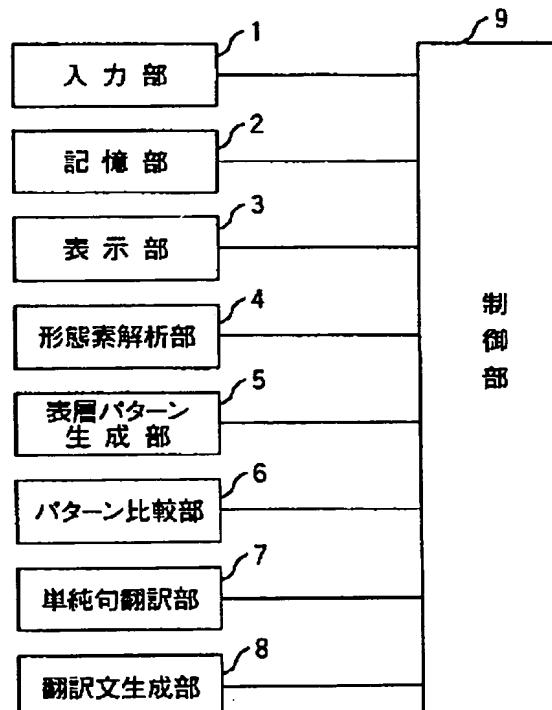
最終頁に続く

(54)【発明の名称】 自然言語の翻訳装置

(57)【要約】

【目的】 文の構造解析や編集をすることなく質の高い訳文を得る。

【構成】 記憶部2には対訳例文データベースを格納する。形態素解析部4は入力部1からの入力文を形態素解析して述語を得る。表層パターン生成部5は形態素解析結果に基づいて入力文の表層パターンを生成する。パターン比較部6は入力文の述語に基づいて対訳例文データベースのインデックスを決定し、更に入力文の表層パターンを用いて当該インデックス下に在る変換パターンおよび対訳例文を検索する。単純句翻訳部7は対訳例文を参照して入力文の表層パターンにおけるカテゴリ・シンボルに対応する単語列を翻訳して変換パターン内の空欄を埋めて目標言語の文字列パターンを得る。翻訳文生成部8は目標言語の文字列パターンに基づいて完全な翻訳文を生成する。こうして、入力文の構造解析や編集をすることなく質の高い訳文を得る。



【特許請求の範囲】

【請求項1】 入力部から入力された自然言語による文章に対して形態素解析部で形態素解析を行い、記憶部に格納されている訳例文データベースから入力文に対応する例文とその対訳との対である対訳例文を対訳例文検索部によって上記形態素解析結果に基づいて検索し、この検索された対訳例文に基づいて翻訳部で入力文章を目標言語に翻訳し、得られた翻訳結果を表示部に表示する自然言語の翻訳装置において、

上記形態素解析部による入力文に対する形態素解析結果に基づいて、上記入力文から、少なくとも用言および付属語の文字列とそれらに前後する単語列の構文カテゴリとによって文の表層的特徴を表した表層パターンを所定の手順で生成する表層パターン生成部を備えると共に、上記対訳例文データベースに蓄積された各対訳例文には、この対訳例文における例文の上記表層パターンを予め付加し、

上記対訳例文検索部は、上記対訳例文データベースから該当する対訳例文を検索するに際して、上記表層パターン生成部によって生成された入力文の表層パターンと上記対訳例文に付加されている例文の表層パターンとの類似度を求ることによって、入力文に類似した例文を有する対訳例文を検索する構成に成したことを特徴とする自然言語の翻訳装置。

【請求項2】 請求項1に記載の自然言語の翻訳装置において、

上記記憶部に格納された対訳例文データベースには、用言の文字列パターンをルートノードとし、当該用言を用いた文から抽出された少なくとも当該用言および付属語の文字列パターンを当該ルートノードから分岐した各ノードとする木構造を成すと共に、上記各ノードの文字列パターンは親ノードの文字列パターンを詳細化した文字列パターンになっているインデックス木を設けて、このインデックス木におけるリーフノードの文字列パターンを上記対訳例文データベースのインデックスとし、上記対訳例文検索部は、入力文から上記形態素解析部での形態素解析結果によって抽出された用言に基づいて、当該用言を表す文字列パターンのルートノードを有するインデックス木を検索し、この検索されたインデックス木を用いて上記対訳例文データベースのインデックスを得る構成に成したことを特徴とする自然言語の翻訳装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 この発明は、自然言語で書かれた文章を自動的に翻訳する自然言語の翻訳装置に関する。

【0002】

【従来の技術】 従来より、機械翻訳装置としては、図7に示すような解析レベルに従って、辞書情報と多数の解析ルールを使用して目標言語との対応が取り易くなるレ

ベルまで原言語による入力文(以下、原文と言う)の解析を行い、原文が表す意味的な内部構造を抽出するという解析プロセスを採用した所謂トランスファー方式によるものが主流である。

【0003】 すなわち、先ず、原文に対する形態素解析によって各単語に対する品詞列を求める。次に、構文解析によって上記品詞列に対する句構造を求める。そして、最後に単語や句の用法に関する種々のデータを使って意味解析を行って依存構造等の最終的な内部構造を得る。こうして、目標言語との対応が取り易くなるレベルまで解析されると、目標言語への変換規則を用いて同レベルの目標言語に変換し、そこから構文生成、形態素生成と生成プロセスを進めて目標言語を生成して行くのである。

【0004】 このように、従来の機械翻訳装置では、解析主導の翻訳プロセスが翻訳処理の前提となっている。ところが、上記従来の解析主導の翻訳システムには以下のようないくつかの欠点がある。

【0005】

(1) 翻訳の専門家ののような柔軟な意訳ができない
目標言語への変換規則は、通常機械的な置き換えによるものであり、分かりやすい表現の訳文にするための知識は反映されてはいない。したがって、得られる訳文は堅い表現になり、非常に分かりにくいものになっている。そのために、現行の機械翻訳装置では、マニュアルによって翻訳結果を“後編集”して分かりやすい訳文に修正したり、マニュアルによって入力文を“前編集”して機械翻訳装置が容易に処理可能な文型に書き換えたりしなければ、妥当な訳文を得ることができないのである。

【0006】 その結果、当然のことながら、人手を介すことなく、翻訳の専門家が訳すようなレベルの“意訳”的訳文を得ることは極めて難しい。

【0007】

(2) 機械システムのメンテナンスや改良が困難
部分的に上記目標言語への変換規則等の解析ルールや経験則を増やして翻訳システムを改良しようとしても、全体の処理アルゴリズムに影響を及ぼしてしまうので変更に伴う負担が大きい。また、翻訳システムを修正できたとしてもヒューリスティックに依存する部分が多く、ヒューリスティックを統一的に制御する有効な手段を備えてはいないために、翻訳改善の対象となった文に対しては良好な改善結果が得られる一方で、別の文章に対しては翻訳精度が低下してしまうという事態が発生し易い。

【0008】 上述のように、解析主導の翻訳システムの欠点を解消すべく、近年、例文主導の翻訳システムが提唱されている。

【0009】 この例文主導の翻訳システムでは、入力文に最も類似した対訳例文(対訳を有する例文)を対訳例文データベースから検索し、この検索した対訳例文の対訳を利用して上記入力文に対する翻訳を得るようにしてい

る。この翻訳システムには、上記対訳例文データベースに対訳例文を追加するだけで性能向上を図れるという利点や、対訳例文によってカバーできる範囲内においては意訳レベルでの翻訳が実施できるという利点がある。

【0010】

【発明が解決しようとする課題】しかしながら、上記従来の例文主導の翻訳システムには以下のような問題点がある。現在提唱されている例文主導の翻訳システムにおいては、上記対訳例文を検索する際に用いるキーとして文章の依存構造を予め用意しておくものが多い。このために、対訳例文の検索に際しては翻訳対象となる入力文の依存構造を求める必要がある。そして、そのためにには、入力文章の形態素解析、構文解析、係り受け解析、意味解析を正確に行わなければならない。

【0011】ところで、一般に、入力文に対する形態素解析および構文解析の際には多数の解析候補が得られる。そして、長く複雑な文章になるほど得られる解析候補の数が増大する。さらに、上記解析候補を絞り込むための意味解析においては、拠り所となる規則が存在しない。そこで、通常は多数の経験則を用意しておいて状況に応じて使い分けることになる。

【0012】その結果、長く複雑な文章になるほど、上記得られた多数の解析候補から文意に沿った候補を一気に絞り込むことが困難になるのである。したがって、上記依存構造を有する対訳例文を上記対訳データベースに格納する上記例文主導の翻訳システムでは、係り受け関係の複雑な文章を正しく翻訳できる確率が低いという問題点がある。

【0013】そこで、この発明の目的は、入力文の構文解析、係り受け解析および意味解析等の解析プロセスを適用する必要がなく且つ“後編集”および“前編集”を実施することなく、係り受けの複雑な入力文であっても質の高い訳文を得ることができる自然言語の翻訳装置を提供することにある。

【0014】

【課題を解決するための手段】上記目的を達成するため、第1の発明の自然言語の翻訳装置は、入力部から入力された自然言語による文章に対して形態素解析部で形態素解析を行い、記憶部に格納されている訳例文データベースから入力文に対応する例文とその対訳との対である対訳例文を対訳例文検索部によって上記形態素解析結果に基づいて検索し、この検索された対訳例文に基づいて翻訳部で入力文章を目標言語に翻訳し、得られた翻訳結果を表示部に表示する自然言語の翻訳装置において、上記形態素解析部による入力文に対する形態素解析結果に基づいて、上記入力文から、少なくとも用言および付属語の文字列とそれらに前後する単語列の構文カテゴリによって文の表層的特徴を表した表層パターンを所定の手順で生成する表層パターン生成部を備えると共に、上記対訳例文データベースに蓄積された各対訳例文にはこ

の対訳例文における例文の上記表層パターンを予め付加し、上記対訳例文検索部は、上記対訳例文データベースから該当する対訳例文を検索するに際して、上記表層パターン生成部によって生成された入力文の表層パターンと上記対訳例文に付加されている例文の表層パターンとの類似度を求めるこによって入力文に類似した例文を有する対訳例文を検索する構成に成したことの特徴としている。

【0015】また、第2の発明は、第1の発明の自然言語の翻訳装置において、上記記憶部に格納された対訳例文データベースには、用言の文字列パターンをルートノードとし、当該用言を用いた文から抽出された少なくとも当該用言および付属語の文字列パターンを当該ルートノードから分岐した各ノードとする木構造を成すと共に、上記各ノードの文字列パターンは親ノードの文字列パターンを詳細化した文字列パターンになっているインデックス木を設けて、このインデックス木におけるリーフノードの文字列パターンを上記対訳例文データベースのインデックスとし、上記対訳例文検索部は、入力文から形態素解析部での形態素解析結果によって抽出された用言に基づいて当該用言を表す文字列パターンのルートノードを有するインデックス木を検索し、この検索されたインデックス木を用いて上記対訳例文データベースのインデックスを得る構成に成したことを特徴としている。

【0016】

【作用】第1の発明では、入力部から入力された自然言語による文章に対して形態素解析部によって形態素解析が実施され、この形態素解析結果に基づいて、表層パターン生成部によって、入力文から、少なくとも用言および付属語の文字列とそれらに前後する単語列の構文カテゴリによって文の表層的特徴を表した表層パターンが所定の手順で生成される。そうすると、対訳例文検索部によって、記憶部の対訳例文データベースに蓄積された各対訳例文に付加されている例文の表層パターンと上記表層パターン生成部で生成された入力文の表層パターンとの類似度が求められる。そして、この類似度に基づいて、入力文に類似した例文を有する対訳例文が検索される。

【0017】以後、この検索された対訳例文に基づいて、翻訳部によって入力文章が目標言語に翻訳され、得られた翻訳結果が表示部に表示される。こうして、文全体の表層的特徴を表す表層パターンを用いた形態素レベルでの類似度算出のみによって、非常に簡単に入力文に対応する対訳例文を検索して質の良い翻訳が得られる。

【0018】また、第2の発明では、入力部から入力された入力文が形態素解析部によって形態素解析され、この形態素解析結果に基づいて入力文の用言が抽出される。そうすると、対訳例文検索部によって、上記抽出された用言を表す文字列パターンのルートノードを有する

インデックス木が検索され、この検索されたインデックス木を用いて上記対訳例文データベースのインデックスが得られる。

【0019】そして、こうして得られたインデックスを用いて、上記対訳例文検索部によって、入力文の表層パターンとの類似度算出の対象となる対訳例文候補が選出される。

【0020】

【実施例】以下、この発明を図示の実施例により詳細に説明する。この発明における自然言語の翻訳装置は、文章の表層パターンを利用して対訳例文を検索する例文主導の翻訳システムを備えた翻訳装置である。

【0021】図1は本実施例の自然言語の翻訳装置における概略プロック図である。以下、便宜上、日本語による原文を英語に翻訳する場合を例に上記自然言語の翻訳装置を説明する。

【0022】入力部1はキーボードや光学文字読み取り装置(OCR)等の入力機器で構成されて、上記対訳例文や翻訳対象の文章等を入力する。記憶部2はRAM(ランダム・アクセス・メモリ)やROM(リード・オンリ・メモリ)等のメモリおよびこのメモリを制御するメモリ制御手段で構成されて、単語辞書や対訳例文データベース等を格納する。表示部3はCRT(カソード・レイ・チューブ)等の表示機器で構成される。

【0023】形態素解析部4は、記憶部2のメモリに格納されている単語辞書を引いて入力文章から単語列を切り出し、品詞列を生成する。さらに、テンスやアスペクト等の情報を得る。表層パターン生成部5は、形態素解析部4による形態素解析結果を用いて、入力部1からの入力文の表層パターンを生成する。パターン比較部6は、後に詳述するようにして、入力文章の表層パターンの候補と記憶部2の上記メモリに格納されている対訳例文データベースに用意されている表層パターンとの比較を行って、入力文章に最も類似した表層パターンを有する対訳例文を検索する。

【0024】単純句翻訳部7は、複雑な埋め込み文のない名詞句(「本」、「その本」、「彼の本」、「美しい本」等)や、空列を含む助動詞列が後続する述語等の単純な語句を対象として、上記単語辞書のような簡単なルールのみに基づいて翻訳処理を実行する。この単純句翻訳部7は、上述した従来型の機械翻訳装置における一部の機能で代用可能であるために、ここでは詳細な説明は省略する。

【0025】翻訳文生成部8は、上記目標言語における単語の並びやテンスおよびアスペクト等の情報をから、目標言語による完全な翻訳文を生成する。尚、この翻訳文生成部8についても従来型の機械翻訳装置における一部の機能で代用可能であるために詳細な説明は省略する。制御部9は、上記入力部1、記憶部2、表示部3、形態素解析部4、表層パターン生成部5、パターン比較部6、単純句翻訳部7および翻訳文生成部8を制御して、入力文

章の翻訳処理を実施する。

【0026】すなわち、上記パターン比較部6で上記対訳例文検索部を構成し、単純句翻訳部7、翻訳文生成部8および制御部9で上記翻訳部を成すのである。

【0027】本実施例の翻訳装置によって実施される翻訳処理の概略は、入力文の表層パターンを用いて入力文に最も類似した対訳例文を上記対訳例文データベースから検索し、検索された対訳例文における対訳を基にして入力文の翻訳文を得る処理である。以下、上記翻訳処理について順を追って詳細に説明する。尚、ここで言う表層パターンとは、文を特徴付ける単語とその他の部分単語列の構文カテゴリによって表されるものである。

【0028】先ず、上記記憶部2のメモリに格納される対訳例文データベースについて説明する。図2および図3は上記対訳例文データベースに関する説明図である。図2は上記対訳例文データベースのインデックス構造を示し、「ある」という動詞が述部となる和文を原文とする複数の対訳例文のインデックス構造を例示している。

【0029】上記インデックスは、述語の終止形「ある」をルートノードとし、その述語「ある」を含む表層の文字列パターン「*は*がある」、「*には*がある」、「*は*に*がある」、…をルートノード以外のノードとする木構造で表現される。尚、上記表層の文字列パターンにおけるパターン要素は、各ノードに存在する述語「ある」に対する必須格、任意格、省略格の格助詞および特徴的な単語である。上記表層の文字列パターンは、リーフノードに行くほど詳細に記述され、子ノードの文字列パターン(例えば、「*は*と*がある」)は親ノードの文字列パターン(例えば、「*は*がある」)を詳細化した文字列パターンになっている。

【0030】そして、上述のようなルートノードを幹とする木構造を有するインデックス木の各リーフノードに係る上記表層の文字列パターンを上記対訳例文データベースのインデックスとし、このインデックスに対訳例文が対応付けられている。したがって、入力文から抽出した述語の終止形をルートノードとするインデックス木を上記入力文の表層の文字列に従って辿って行くことによって、該当する対訳例文を検索するためのインデックスを決定できるのである。

【0031】図3は、「*には*がある」という表層の文字列パターンを有する和文を原文とする対訳例文を蓄積した対訳例文データベースの構造例を示す。図3に示すように、上記対訳例文データベースは、上記インデックス(上記インデックス木のリーフノードに係る表層の文字列パターン)、表層パターン、変換パターンおよび対訳例文からなる層構造を成している。

【0032】ここで、上記表層パターンは本実施例の中心となるデータ構造であり、上述したように文を特徴付ける単語(以下、特徴単語と言う)とその他の部分単語列の構文カテゴリによって表される。ここで、上記特徴単

語とは、動詞、助詞および一部の特徴的な名詞であり、図2に示すインデックスにおける各ノードの文字列パターンに具体的に表記された単語に対応する。また、上記構文カテゴリとは、上記特徴単語に前後する単語列(すなわち、上記インデックスでは“*”に対応する部分単語列)の簡単な句構造を表すものである。

【0033】次に、上記表層パターンの構成法について説明する。

構文カテゴリ	カテゴリ・シンボル
単純名詞句	N
埋め込み文によって装飾された名詞句	VP・N
動詞句	VP
...	...

【0034】上述のようにして構成される表層パターンを用いて、上記対訳例文データベースは次のように構成される。以下、図3に従って対訳例文データベースの構成について具体的に説明する。

【0035】“*には*がある”というインデックス下には、次のようなパターン1～パターン3と命名された3つの表層パターンが存在する。すなわち、

パターン1 = “N1にはN2がある”

= “単純名詞句1+「には」+単純名詞句2+「が」+「ある」”

パターン2 = “VP・NIにはN2がある”

= “連体修飾述句+単純名詞句1+「には」+単純名詞句2+「が」+「ある」”

パターン3 = “VPにはNがある”

= “述句+「には」+単純名詞句+「が」+「ある」”

【0036】さらに、各表層パターン下には、その表層パターンを有する和文を英訳する際に用いられる変換パターンが存在する。

例えば、パターン1 = “N1にはN2がある”

= “単純名詞句1+「には」+単純名詞句2+「が」+「ある」”

に対しては、

変換パターン = “There BE T(N2) in T(N1).”

が対応付けられており、

“There”+BE動詞+単純名詞句2の翻訳結果+“in”+単純名詞句1の翻訳結果

が変換されるべき英文のパターンであることを示している。

【0037】ここで、上記変換パターンに見られる“T(x)”という表記は、句“x”に対応する単語列を上記単純句翻訳部7(図1参照)によって翻訳した結果を表す。例えば、CASE01に示す対訳例文の場合には、“x”は「庭」を表す単純名詞句であり、“T(x)”は“garden”である。また、“Tch(x)”という表記は、CASE番号“h”を有する対訳例文の対訳英文を表す。例えば、CASE11に示す対訳例文の場合には、“x”は「彼が学会誌に発表した論文」を表す埋め込み文を含む名詞句であり、CASE番号

(1) 対象となる文の中心用言とそれに係る任意格を含めた格助詞、接続助詞および特徴的な名詞とを夫々抽出して上記特徴単語とする。

(2) (1)で抽出された特徴単語に前後する部分単語列の上記構文カテゴリを設定する。そして、その設定された構文カテゴリを次のようにカテゴリ・シンボルに置き換える。

構文カテゴリ	カテゴリ・シンボル
単純名詞句	N
埋め込み文によって装飾された名詞句	VP・N
動詞句	VP
...	...

“11”の対訳例文に記載された同じ和文に対する対訳英文を取り出すことによって、“Tc11(x)”=“the paper which he published in a scholar journal”が得られる。尚、上記CASExxは、具体的な例文と対訳との対から成る対訳例文を表す。例えば、CASE01の場合には、和文「庭には池がある」と対を成す英訳文は“*There is a pond in the garden*”である。

【0038】つまり、上記変換パターンは一種のテンプレートとなっており、対応する表層パターンを構成する上記特徴単語に前後する部分単語列の翻訳結果で上記テンプレートの空欄を埋めることによって翻訳英文が得られるのである。

【0039】上述のよう構造を有する対訳例文データベースとして大量の対訳例文を蓄積しておけば、入力文の表層パターンと類似若しくは一致した表層パターンを有する対訳例文を対訳例文データベースから検索することによって、質の高い翻訳文を得ることが容易に可能となるのである。

【0040】ここで、上述のよう表層パターンを用いて翻訳を実施することによって、次のような利点が得られるのである。

【0041】(A)上記対訳例文データベースから入力文に類似若しくは一致する対訳例文を検索する際に実施される表層パターンのマッチングは、1次元的な形態素解析レベルでのパターンマッチングである。したがって、依存構造解析のように2次元的な解析を行う必要がない。具体的には、上記依存構造解析の場合には、入力文全体に対する係り受け解析および意味処理を含めた構文解析を必要とする。これに対して、表層パターンのマッチングの場合には、文字列のパターンマッチング、形態素解析および品詞列に対する極簡単なパターン認識処理しか必要とはしない。したがって、入力文章の解析処理が非常に単純なものとなる。

【0042】このように、上記対訳例文の検索に伴う解析処理が簡単になることによって、従来型の例文主導の翻訳システムに比較して長く複雑な入力文章に対する翻訳処理時間が大幅に短縮される。

【0043】(B)上記従来型の例文主導による翻訳システムで実施される依存構造解析は、局所的に解析ルールを適用してマッチングを行い、得られた結果を積み上げるボトムアップ方式である。そのために、部分的には正しく構造が解析されているにも拘わらず、文章全体としては係り受け関係や句のまとまりが誤っている解析候補が生成される場合が多い。

【0044】これに対して、上記表層パターンは文全体を規定したものであるために、表層パターンのマッチング処理に際しては巨視的に見た場合の翻訳の失敗を避けることができる。また、その結果、訳文候補の組み合せの爆発を避けることができる。以上の理由から、本実施例における表層パターンを用いた翻訳システムによれば、長く複雑な文章に対する翻訳の精度が飛躍的に向上するのである。

【0045】次に、上記入力部1から入力された入力文章から上記表層パターンを抽出し、記憶部2のメモリに格納された対訳例文データベースから上記入力文章に類似した対訳例文を上記抽出された入力文章の表層パターンに基づいて検索する対訳例文検索処理動作について説明する。

【0046】図4および図5は、上記制御部9によって記憶部2、形態素解析部4、表層パターン生成部5およびパターン比較部6を制御して実施される対訳例文検索処理動作のフローチャートである。以下、図4に従って、上記対訳例文検索処理動作について詳細に説明する。

【0047】ステップS1で、上記形態素解析部4によって、入力部1から入力された入力文“S”的形態素が解析されて単語列および品詞列が切り出され、テンスおよびアスペクト等の情報が得られる。そして、得られた入力文Sの単語列および品詞列から入力文Sの述語

“V”が決定される。ステップS2で、上記パターン比較部6によって、上記ステップS1において決定された述語Vをキーワードとして、図2の構造を有して上記対訳例文に関連付けられた複数のインデックス木から当該述語Vと同じ文字列パターンをルートノード(以下、“ルートノードV”と言う)とするインデックス木が検索される。

【0048】ステップS3で、さらに上記パターン比較部6によって、上記検索されたインデックス木におけるルートノードVから分岐している各子ノードchild(V)の文字列パターンのパターン要素をキーワードとして、全子ノードchild(V)の文字列パターンと入力文Sの文字列とが比較される。ステップS4で、上記キーワードであるパターン要素が入力文Sの文字列中に在るような子ノードchild(V)が存在するか否かが判別される。その結果存在すればステップS5に進み、存在しなければ上記対訳例文データベース内に入力文Sに類似する対訳例文はないとして対訳例文検索処理動作を終了する。

【0049】ステップS5で、当該子ノードを親ノード

“F”とする。ステップS6で、上記パターン比較部6によって、子ノードchild(F)に係る上記パターン要素をキーワードとして、全子ノードchild(F)の文字列パターンと入力文Sの文字列とが比較される。ステップS7で、上記パターン要素が入力文Sの文字列中に在るような子ノードchild(F)が存在するか否かが判別される。その結果、存在すればステップS5に戻って当該子ノードchild(F)から分岐したノードに対する処理に移行する。一方、存在しなければステップS8に進む。

【0050】ステップS8で、上記ノードFはリーフノードであるから、このノードFの文字列パターンが入力文Sに類似した対訳例文を検索する際のインデックスであると決定される。ここで、便宜上、上記インデックスを“*P₁*P₂*…*P_j*…*P_J·V”と表す。但し、“P_j(j=1~J)”はj番目のインデックス要素であり、“*”は上記インデックス要素に前後する部分文字列である。ステップS9で、上記ステップS8において決定されたインデックスの文字列パターンにおけるインデックス要素が参照されて、入力文Sの文字列が上記インデックス要素と同じ文字の箇所で分割される。その際に、上記入力文Sの文字列に上記インデックス要素と同一の部分文字列が複数あるために分割箇所が一意に決まらない場合には、総ての分割候補が求められて保持される。ここで、上記分割候補がI個あるとした場合には、このI個の分割候補の集合{b_i}は次のように表される。

{b_i}={conc(S_{ij}·P_j)j=1~J}i=1~I

但し、S_{ij}: j番目のインデックス要素P_jの直前に位置する“*”に対応する部分文字列

【0051】ステップS10で、分割候補番号iと表層パターン番号kとに“1”がセットされる。また、マッチング評価値E_kと最大マッチング評価値E_{k'}と最大マッチング評価値を呈する表層パターン番号k'’と最大評価値を呈する分割候補番号i'に“0”がセットされる。ステップS11で、上記表層パターン生成部5によって、i番目の分割候補b_iの各部分文字列(S_{ij})_{j=1~J}に対して形態素解析が実施されて、以下のような分割候補b_iの表層パターンbp_iが求められる。

bp_i=[X_{ij}·P_j]j=1~J

但し、X_{ij}: 部分文字列S_{ij}を形態素解析して得られた品詞列H₁, …, H_r, …, H_Rに対して割り当てられる上記カテゴリ・シンボル列

【0052】上記カテゴリ・シンボル列X_{ij}の割り当ては、次のような割り当てるルールを適用して実施される。
(a) 品詞H_Rが動詞、動詞に続く付属語、名詞に続く述語型助動詞である場合にはカテゴリ・シンボル“VP”を割り当てる。

(b) 品詞H_Rが名詞、名詞に続く接辞であり、且つ、r < Rであるrに対して連体形の動詞である品詞H_rが存在する場合には、カテゴリ・シンボル列“VP·N”を割

り当てる。

(c) 品詞 H_R が名詞、名詞に続く接辞であり、且つ、 $r < R$ である r に対して動詞である品詞 H_r が存在しない場合にはカテゴリ・シンボル “N” を割り当てる。

【0053】ステップS12で、上記対訳例文データベースから上記ステップS8において決定されたインデックス下に在る k 番目の表層パターン(以下、任意のインデックス下に在る表層パターンをインデックス内表層パターンと言う) dp_k が読み出される。ここで、当該インデックス下には K 個のインデックス内表層パターン dp_k が在るものとすると、この K 個のインデックス内表層パターンの集合{ dp_k }は次のように表される。

$$\{dp_k\} = \{[C_{kj} \cdot P_j]_{j=1 \sim J}\}_{k=1 \sim K}$$

但し、 C_{kj} : j 番目の上記特徴単語 P_j の直前に位置するカテゴリ・シンボル列

つまり、上記インデックス内表層パターンは、入力文 S と同じ述語 V を含む入力文 S と同じ上記表層の文字列パターンを有する表層パターンであると言える。ステップS13で、上記パターン比較部6によって、上記ステップS11において求められた入力文 S の表層パターン bp_i のカテゴリ・シンボル列 X_{ij} と上記ステップS12において読み出されたインデックス内表層パターン dp_k のカテゴリ・シンボル列 C_{kj} とが、総ての j について比較される。その結果、 $X_{ij} = C_{kj}$ または $X_{ij} \neq C_{kj}$ であればステップS18に進む。一方、 $X_{ij} \neq C_{kj}$ であればステップS14に進む。ここで、上記 “ $X_{ij} \neq C_{kj}$ ” とは、カテゴリ・シンボル X_{ij} あるいはカテゴリ・シンボル C_{kj} のうち何れか一方のヘッドフィーチャーが他方のカテゴリ・シンボルと一致する場合である。

【0054】ステップS14で、上記インデックス内表層パターン dp_k の表層パターン番号 k の内容が最大値 “K” より小さいか否かが判別される。その結果最大値 “K” より小さければステップS15に進み、そうでなければステップS16に進む。ステップS15で、表層パターン番号 k の内容がインクリメントされてステップS12に戻り、次のインデックス内表層パターンの処理に移行する。ステップS16で、上記分割候補番号 i の内容が最大値 “I” より小さいか否かが判別される。その結果最大値 “I” より小さければステップS17に進み、そうでなければステップS21に進む。ステップS17で、分割候補番号 i の内容がインクリメントされてステップS11に戻り、入力文 S の次の分割候補の表層パターンに対する処理に移行する。

【0055】ステップS18で、上記ステップS13での比較結果に基づいて、入力文 S (分割候補 b_i) の表層パターン bp_i とインデックス内表層パターン dp_k との間のマッチング評価値 E_k が以下のようにして算出される。すなわち、先ず、上記分割候補 b_i の表層パターン bp_i のカテゴリ・シンボル列 X_{ij} と上記インデックス内表層パターン dp_k のカテゴリ・シンボル列 C_{kj} との比較結果に基づ

いて、以下のようにマッチ度 $C E_{kj}$ が設定される。

【0056】上記マッチ度 $C E_{kj}$ は次のように設定される。

(イ) カテゴリ・シンボル列 X_{ij} とカテゴリ・シンボル列 C_{kj} とが完全に一致する場合 ($X_{ij} = C_{kj}$)

例えば、 X_{ij} 及び C_{kj} が共に埋め込み文によって装飾された名詞句 “VP・N” である場合には、マッチ度 $C E_{kj}$ に “1.0” を与える。

(ロ) カテゴリ・シンボル列 X_{ij} とカテゴリ・シンボル列 C_{kj} のヘッドフィーチャーとが一致する場合 ($X_{ij} \neq C_{kj}$)

例えば、 X_{ij} が単純名詞句 “N” で C_{kj} が埋め込み文によって装飾された名詞句 “VP・N” である場合には、マッチ度 $C E_{kj}$ に “0.5” を与える。

(ハ) カテゴリ・シンボル列 X_{ij} のヘッドフィーチャーとカテゴリ・シンボル列 C_{kj} とが一致する場合 ($X_{ij} \neq C_{kj}$)

例えば、 X_{ij} が埋め込み文によって装飾された名詞句 “VP・N” で C_{kj} が単純名詞句 “N” である場合には、マッチ度 $C E_{kj}$ に “0.5” を与える。

【0057】こうして、総ての “j” についてマッチ度 $C E_{kj}$ が与えられると(すなわち、分割候補 b_i の表層パターン bp_i とインデックス内表層パターン dp_k とが一致あるいは類似すると)、 j 個のマッチ度 $C E_{kj}$ の和が算出されて表層パターン bp_i とインデックス内表層パターン dp_k との間のマッチング評価値 E_k が得られる。

【0058】ステップS19で、上記記憶部2のメモリに現在保持されている最大マッチング評価値 E_k' と上記算出されたマッチング評価値 E_k とが比較される。その結果、当該マッチング評価値 E_k の方が最大マッチング評価値 E_k' よりも大きい場合にはステップS20に進む。一方、最大マッチング評価値 E_k' 以下であればステップS14に戻って、次のインデックス内表層パターンが在れば次のインデックス内表層パターンに対する処理に移行する。ステップS20で、上記記憶部2によって、メモリに格納されている上記最大マッチング評価値 E_k' を呈する表層パターン番号 k' が当該表層パターン番号 “k” に更新され、最大マッチング評価値 E_k' を呈する分割候補番号 i' が当該分割候補番号 “i” に更新され、そして最大マッチング評価値 E_k' が当該マッチング評価値 “ E_k ” に更新される。その後、ステップS14に戻って、次のインデックス内表層パターンが在れば次のインデックス内表層パターンに対する処理に移行する。

【0059】ステップS21で、入力文 S に係る総ての分割候補 b_i ($i = 1 \sim N$) および上記対訳例文データベースにおける当該インデックス下に在る総てのインデックス内表層パターン dp_k ($k = 1 \sim K$) に関する検索処理が終了したので、最大マッチング評価値 E_k' を呈するインデックス内表層パターン下に在る対訳例文が output される。また、最大マッチング評価値 E_k' を呈する分割候補

の表層パターンが出力される。こうして、入力文Sの表層パターンに類似したあるいは一致した表層パターンを有する対訳例文が出力されて、対訳例文検索処理動作を終了する。

【0060】このようにして、入力文Sに類似あるいは一致した対訳例文が得られると、当該対訳例文と当該対訳例文上に在る上記変換パターンとを入力文Sに適用して目標言語の具体化された文字列パターンを得る。その際ににおける入力文Sへの適用とは、当該変換パターン内における表記T(x)に対応する当該分割候補b₁内における部分文字列S_{1j}の上記単純句翻訳部7による翻訳や、当該変換パターン内における表記“T_{ch}(x)”で指定された対訳例文を用いた部分翻訳を意味する。

【0061】こうして、上記目標言語の具体化された文字パターンが得られると、上記翻訳文生成部8によって、形態素解析部4による形態素解析で得られたテンスおよびアスペクトに関する情報や訳文生成ルールに基づいて、目標言語に具体化された文字パターンの時制、人称および数等の表現の検査/修正が行われて完全な翻訳文が生成される。そして、生成された翻訳結果は表示部3に出力されて表示される。

【0062】次に、本実施例における翻訳装置によって実施される例文主導の翻訳処理について、入力例文を上げて図1～図5を参照して順を追って具体的に説明する。

【0063】和文による入力文S「彼が買った本には落丁があった」が入力部1から入力される。そうすると、形態素解析部4で形態素解析が行われて述語V「ある」が決定され、入力文Sの時制情報“過去”が得られる。…ステップS1

上記述語V「ある」がルートノードになっている図2に示すインデックス木が検索される。そして、この検索されたインデックス木の子ノードの文字列パターンと入力文S「彼が買った本には落丁があった」の文字列とが比較されて、インデックス「*には*がある」が決定される。…ステップS2～ステップS8

【0064】上記インデックス「*には(P₁)*が(P₂)ある」が参照されて、入力文S「彼が買った本には落丁があった」が分割される。その際に、インデックス要素P₁(=「には」)とこれに続くP₂(=「が」)とは入力文S中に各々一回しか出現しないので、分割候補はb₁唯一つだけ存在する。

b₁='彼が買った本(S₁₁)/には(P₁)/落丁(S₁₂)/が(P₂)/ある(V)…ステップS9

次に、上記分割候補b₁内の部分文字列S₁₁(='彼が買った本')及び部分文字列S₁₂(='落丁')に対する形態素解析が実施される。そして、上記部分文字列S₁₁(='彼が買った本')には上記割り当てルール(b)が適用されてカテゴリ・シンボルX₁₁(=VP・N)に変換される。一方、部分文字列S₁₂(='落丁')には割り当てルール(c)が適

用されてカテゴリ・シンボルX₁₂(=N)に変換される。その結果、上記入力文Sにおける分割候補b₁の表層パターンbp₁が次のように求められる。

bp₁="VP・NにはNがある"…ステップS10,ステップS11

【0065】上記対訳例文データベースにおけるインデックス「*には*がある」下にはパターン1,パターン2およびパターン3と命名された3つのインデックス内表層パターンdp₁,dp₂,dp₃が存在する。そこで、上記分割候補b₁の表層パターンbp₁と各インデックス内表層パターンdp₁,dp₂,dp₃の夫々とが比較される。

① bp₁とdp₁との比較

bp₁="VP・N(X₁₁)にはN(X₁₂)がある"

dp₁="N1(C₁₁)にはN2(C₁₂)がある"

したがって、X₁₁=C₁₁ → マッチ度CE₁₁=0.5

X₁₂=C₁₂ → マッチ度CE₁₂=1.0

マッチング評価値E1=1.5

② bp₁とdp₂との比較

bp₁="VP・N(X₁₁)にはN(X₁₂)がある"

dp₂="VP・N1(C₂₁)にはN2(C₂₂)がある"

したがって、X₁₁=C₂₁ → マッチ度CE₂₁=1.0

X₁₂=C₂₂ → マッチ度CE₂₂=1.0

マッチング評価値E2=2.0

③ bp₁とdp₃との比較

bp₁="VP・N(X₁₁)にはN(X₁₂)がある"

dp₃="VP(C₃₁)にはN(C₃₂)がある"

したがって、X₁₁≠C₃₁

分割候補b₁の表層パターンbp₁と見出し内表層パターンdp₃とは別表層パターンである。…ステップS12～ステップS18

【0066】①および②での比較結果により、マッチング評価値E1(=1.5) < マッチング評価値E2(=2.0)

であるから、インデックス内表層パターンdp₂が入力文Sに最も類似したインデックス内表層パターンであると確定される。その結果、類似対訳例文としてCASE11と命名された

例文「彼が学会誌に発表した論文には誤りがある」

対訳「There are some errors in the paper

which he published in a scholar journal」

の対がOutputされる。さらに、上記入力文Sの表層パターン

bp₁="VP・NにはNがある"がOutputされる。

…ステップS19～ステップS21

【0067】こうして、入力文S「彼が買った本には落丁があった」の類似対訳例文が得られると、この得られた対訳例文上に在る上記変換パターン

「There BE T(N2) in T_{ch}(VP・N1).」

に入力文Sが次のように適用される。

T_{ch}(VP・N1=彼が買った本) → 「the book which

he bought」

T(N2=落丁) → 「missing page」

但し、この場合には、変換パターン「There BE T(N2) in Tc12(VP·N1).」下には、例えば、

CASE12 VP=彼が買った

N1=本 N2=誤り

There are some errors in the book which he bought.

なる対訳例文が記述されているものとする。

【0068】こうして、和文による上記入力文S「彼が買った本には落丁があった」の目標言語(英語)に具体化された次のような文字列パターン記述が得られる。

「There BE missing page in the book which he bought.」

以後、この目標言語に具体化された文字列パターンと上記時制情報とに基づいて、上記訳文生成ルールを適用して、目標言語による翻訳文

「There were some missing pages in the book which he bought.」

を得るのである。

【0069】上述の例では、説明の便宜を図るためにごく簡単な係り受け構造しか持たないような入力文Sの翻訳プロセスについて述べているが、更に複雑な係り受け構造を有する文章に対しても適切な翻訳文を得ることが可能である。例えば、以下のようないい入力文

「ハードウェアの構成は、本体とKBD, FDが一体になっているスタンダードアロン型と、本体と一部が分離しているデスクトップ型の2種類があります。」

は、並列句が多く係り受け関係が複雑である。したがって、入力文章を一から解析する従来の解析主導の翻訳システムや依存構造を用いた例文主導の翻訳システムでは、入力文の解析段階で正しい解析結果を得ることが極めて困難である。したがって、高い翻訳精度は得られず、翻訳の専門家のような意訳ができず翻訳の質は低い。

【0070】ところが、本実施例によれば、以下のように高精度で且つ質の高い翻訳文が得られるのである。すなわち、図6に示すように、上記対訳例文データベースに、

- ・インデックス

- “*は*と*の2*がある”

- ・インデックス内表層パターン

- “N1はVP1·N2とVP2·N3の2N4がある”

- ・変換パターン

「There are two N4 of N1 : N2 and N3.

In the N2, Tc(VP1).

In the N3, Tc(VP2).」

- ・対訳例文

N1=推論の方式 N2=帰納法 N3=演繹法 N4=種類

VP1=事実から規則を導く VP2=規則から事実

を導く

「There are two kinds of inference method : induction and deduction.

In the induction, rules are inferred from facts.

In the deduction, facts are inferred from rules.」

を格納しておく。

【0071】上記入力部1から上記入力文S「ハードウェアの構成は、本体とKBD, FDが一体になっているスタンダードアロン型と、本体と一部が分離しているデスクトップ型の2種類があります。」が入力されると、上述のように形態素解析部2によって述語V「ある」が決定される。そして、上記パターン比較部5によってルートノードV “ある”的インデックス木が検索され、入力文Sの文字列に対応するインデックス “*は*と*の2*がある”が求められる。こうして、上記対訳例文データベースのインデックスが決定されると、上述と同様に、決定されたインデックス下に在るインデックス内表層パターン、変換パターンおよび対訳例文を用いて入力文Sの目標言語に具体化された文字列パターン記述が得られるのである。

【0072】このように、長く複雑な係り受けを有する入力文章であっても、その入力文章の表層パターンと同じ表層パターンを呈する対訳例文を対訳例文データベースに登録しておくだけで、翻訳生成に失敗することはないのである。また、長い文章の場合には、文意を取り易いように変換パターンおよび対訳例文の対訳を夫々複数に分割して(図6の場合には3つに分割)意訳するパターンで記述しておくことによって、専門家による翻訳に近い意訳が可能となる。

【0073】上述のように、本実施例では、入力文章の表層の文字列パターンのマッチングおよび入力文章の文字列における上記特徴単語に前後する部分単語列の上記構文カテゴリのマッチングのみを実施すればよく、入力文章を解析して得られた複雑な依存構造によるマッチングを実施する必要がない。したがって、任意格や並列句を含む複雑な係り受け構造を有する入力文章にも容易に対処できる。

【0074】尚、本実施例の翻訳装置では上記対訳例文をどれだけ網羅するかによって翻訳性能が決まる。一方、文の表層の文字列パターンの木を使用して対訳例文データベースをインデキシングするようにしている。したがって、本実施例の翻訳装置によれば、文法の専門家でなくとも系統的に対訳例文を増やして行くことが可能であり、翻訳性能の向上や翻訳システムの改良やメンテナンスを容易に実施できる。

【0075】この発明における対訳例文検索処理動作のアルゴリズムは図4および図5に示すフローチャートに限定されるものではない。また、上記対訳例文データベースの具体的構成は、図3および図6に示すような構成

に限定されるものではない。

【0076】

【発明の効果】以上より明らかのように、第1の発明の自然言語の翻訳装置は、形態素解析部による入力文の形態素解析結果に基づいて、上記入力文から、少なくとも用言および付属語の文字列とそれらに前後する単語列の構文カテゴリとによって文の表層的特徴を表す表層パターンを表層パターン生成部によって生成し、対訳例文検索部によって、上記表層パターン生成部で生成された入力文の表層パターンと上記対訳例文に付加された例文の表層パターンとの類似度を求めるこによって入力文に類似した例文を有する対訳例文を検索するようにしたので、形態素レベルでの類似度算出によって上記対訳例文データベースから容易に該当する対訳例文を検索し、この検索された対訳例文を用いて例文主導の翻訳処理を実施できる。

【0077】したがって、この発明によれば、入力文の構文解析、係り受け解析および意味解析等の2次元的な解析プロセスを適用することなく、“後編集”および“前編集”の実施の必要のない例文主導の翻訳処理を非常に簡単に且つ短時間に実施できるのである。

【0078】さらに、その際における上記対訳例文検索部による類似度算出は、文全体の表層的特徴を表した表層パターンを用いて実施される。したがって、この発明によれば、係り受けの複雑な入力文であっても質の高い訳文を容易に得ることができる。

【0079】また、第2の発明の自然言語の翻訳装置は、記憶部に格納された対訳例文データベースに、用言の文字列パターンをルートノードとし、当該用言を用いた文から抽出された少なくとも当該用言および付属語の文字列パターンを各ノードとする木構造を有するインデックス木を設けて、このインデックス木のリーフノードの文字列パターンを上記対訳例文データベースのインデックスとし、上記対訳例文検索部は、上記形態素解析部

での形態素解析結果によって抽出された用言に基づいてインデックス木を用いて上記対訳例文データベースのインデックスを得るような構成にしたので、得られたインデックスに基づいて、上記対訳例文検索部による類似度計算の対象となる対訳例文候補を容易に選出できる。

【0080】したがって、上記対訳例文検索部は、上記インデックスに基づいて選出された対訳例文候補に付加されている上記表層パターンに付いてのみ上記入力文の表層パターンとの類似度を求めればよく、入力文に類似した例文を有する対訳例文の検索を更に容易に且つ短時間に実施できる。

【図面の簡単な説明】

【図1】この発明の自然言語の翻訳装置におけるプロック図である。

【図2】図1における記憶部に格納された対訳例文データベースを検索する際に使用されるインデックス木の説明図である。

【図3】対訳例文データベースの構成例を示す図である。

【図4】対訳例文検索処理動作のフローチャートである。

【図5】図4に続く対訳例文検索処理動作のフローチャートである。

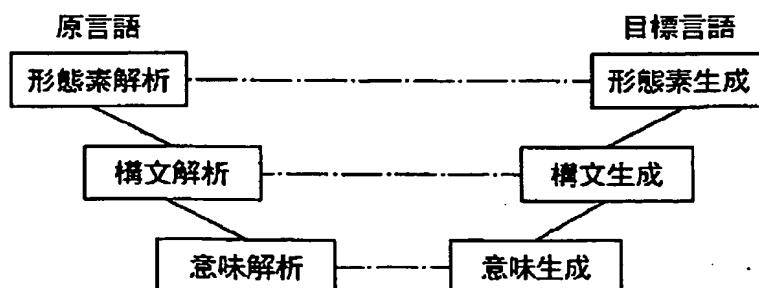
【図6】図3とは異なる対訳例文データベースの構成例を示す図である。

【図7】解析主導の翻訳プロセスにおける解析レベルの説明図である。

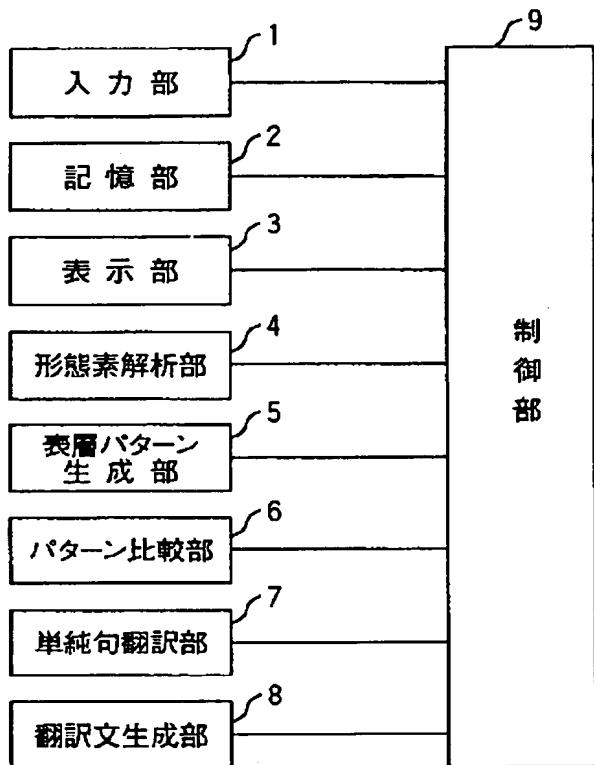
【符号の説明】

1…入力部、	2…記憶部、 3…
表示部、	4…形態素解析部、 5
…表層パターン生成部、	6…パターン比較
部、 7…単純句翻訳部、	8…翻訳文生成部、 9…制御部。

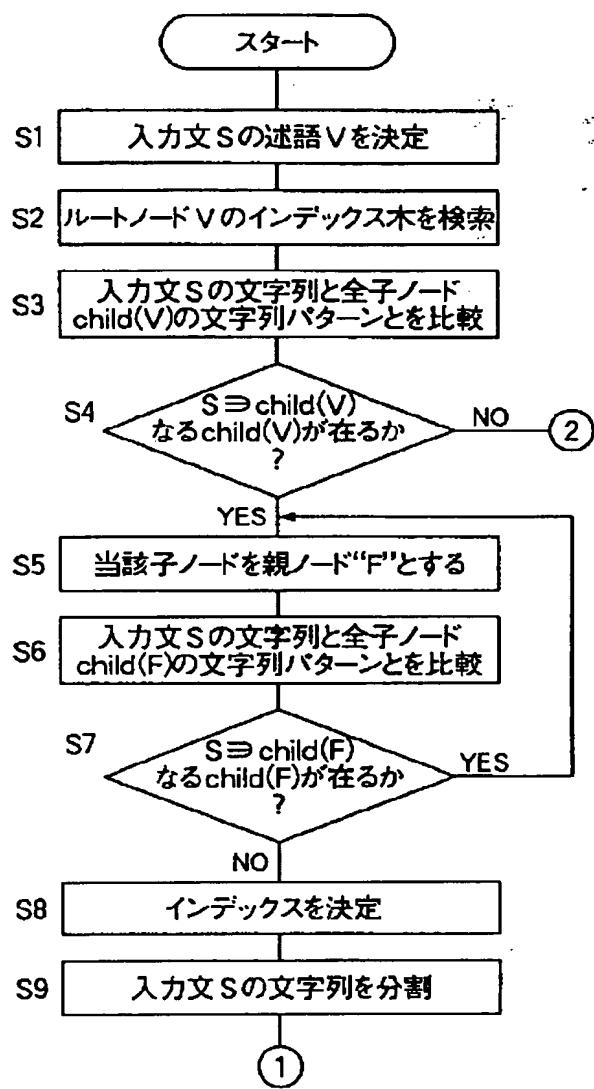
【図7】



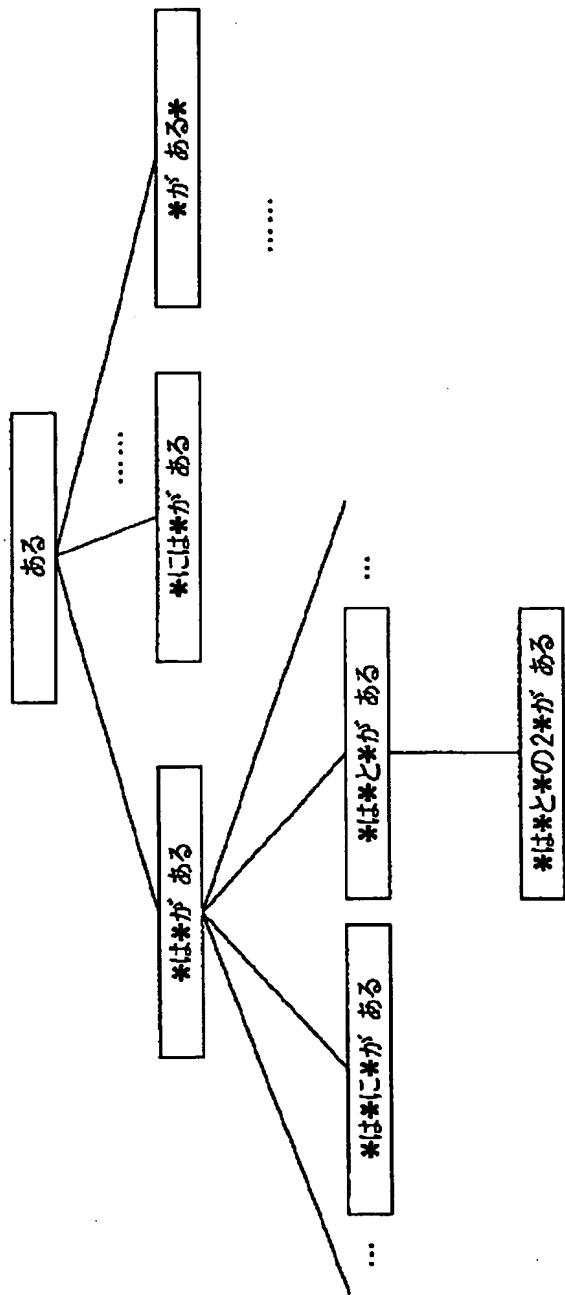
【図1】



【図4】



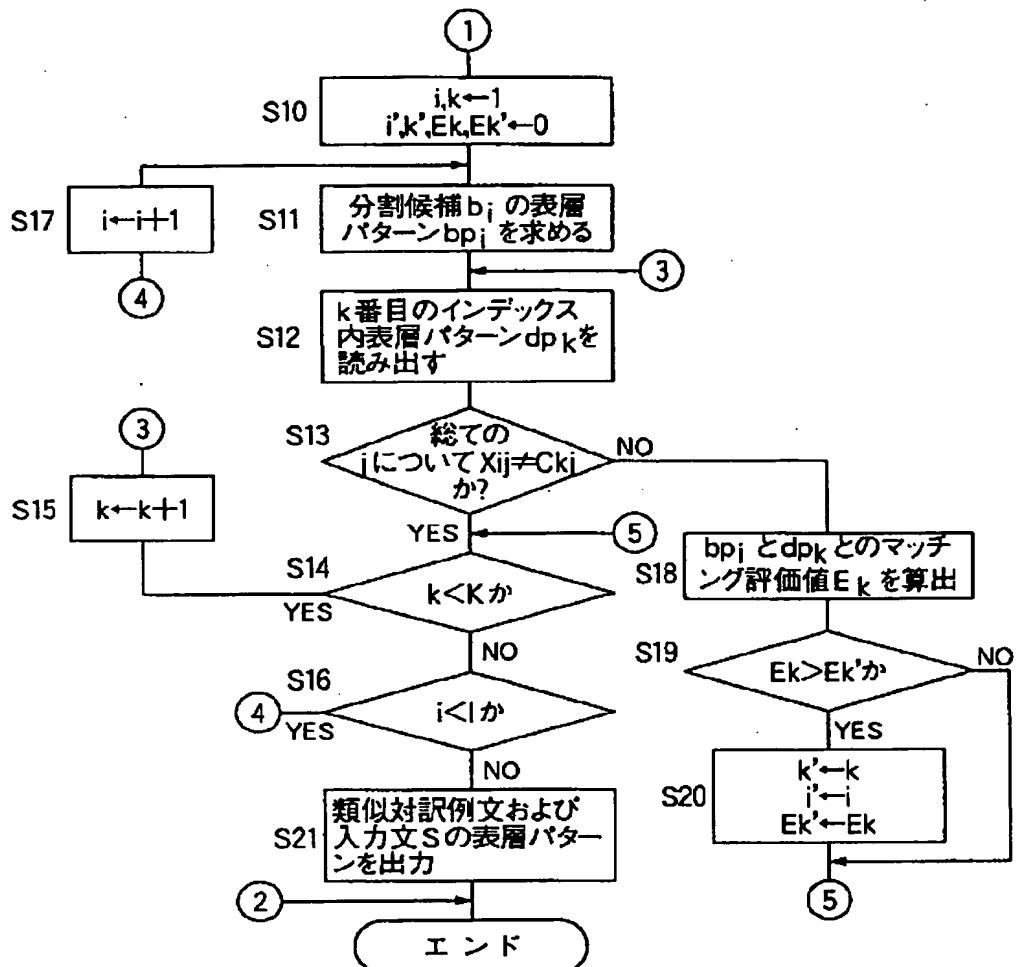
【図2】



【図3】

*には*がある	
パターン1 N1にはN2がある	
	There BE T(N2) in T(N1).
	CASE01 N1=庭 N2=池 There is a pond in the garden.
CASE02 N1=その部屋 N2=書棚 There is a bookshelf in the room.	
パターン2 VP・N1にはN2がある	
	There BE T(N2) in Tc ₁₁ (VP・N1).
	CASE11 VP=彼が学会誌に発表した N1=論文 N2=誤り There are some errors in the paper which he published in a scholar journal
パターン3 VPにはNがある	
	There BE T(N) in Tc ₂₁ (VP).
	CASE21 N=大きな問題 VP=それを実現する There is a big problems in realizing it.
対訳例文	
インデックス	変換パターン
インデックス内 表層パターン	

【図5】



【図6】

<p>*は*と*の2*がある パターン1 N1はVP1・N2とVP2・N3の2N4がある</p>	<p>There are two N4 of N1 : N2 and N3. In the N2, Tc(VP1). In the N3, Tc(VP2).</p>		
	<p>CASE01 N1=推論の方式 VP1=事実から規則を導く</p>	<p>N2=帰納法 VP2=規則から事実を導く</p>	<p>N3=演绎法 VP3=規則から事実を導く</p>
	<p>There are two kinds of inference method : induction and deduction. In the induction, rules are inferred from facts. In the deduction, facts are inferred from rules.</p>		
			<p>対訳例文 インデックス インデックス内 変換パターン 表層パターン</p>

フロントページの続き

(72)発明者 小渕 保司
大阪府大阪市阿倍野区長池町22番22号 シ
ヤープ株式会社内